Lead Scoring Case Study

1: Aditya Dhir

2: Samir Kshirsagar

3: Sharma Nitinkumar Rajesh

Introduction

- In the fiercely competitive landscape of online education, the ability to convert leads into paying customers is a game-changer. X Education, a key player in the industry, faces a challenge – despite a steady influx of leads, their conversion rate remains below par. In this blog, we embark on a journey to build a logistic regression model that assigns lead scores, revolutionizing the way X Education identifies and pursues potential customers.

Problem Statement

X Education aims to boost its lead conversion rate from 30% to an ambitious 80%. The objective is to construct a model capable of assigning lead scores between 0 and 100, empowering the sales team to prioritize interactions with high-potential leads.

Data Exploration

- ‘Leads.csv' comprises data on 9240 leads with 37 columns.
- Seven columns are numeric, and 30 are non-numeric or categorical.
- The current conversion rate stands at 39%.
- Attributes: Various, including Lead Source, Total Time Spent on Website, Total Visits, and Last Activity.
- Target Variable: 'Converted' (1 for conversion, 0 for non-conversion)
- Challenge: Handling 'Select' level in categorical variables.

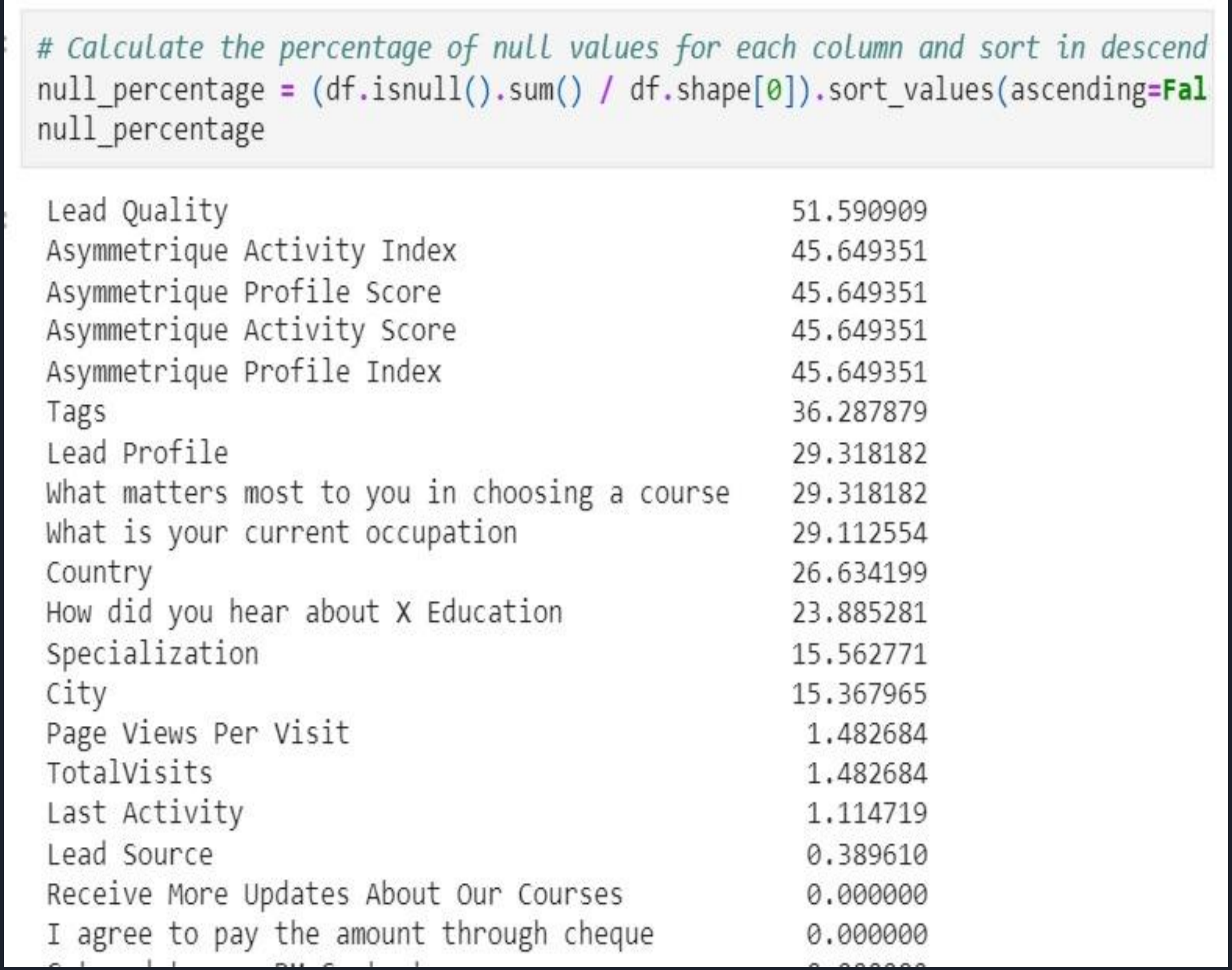
Goals

1. **Logistic Regression Model:** Develop a model for lead scoring.
2. **Adaptability:** Address future company requirements, solving specific problems outlined.
3. **Documentation and Presentation:** Deliver a comprehensive Jupyter notebook, a Word document answering company queries, and a concise presentation summarizing the analysis.

Analysis Approach

1. **Data Exploration and Preprocessing:** Understand and clean the dataset.
2. **Feature Engineering:** Identify influential features affecting lead conversion.
3. **Model Building:** Create a logistic regression model for lead scoring.
4. **Evaluation Metrics:** Assess model performance using relevant metrics.
5. **Adaptability to Future Changes:** Tackle additional company problems, ensuring model flexibility.

Data Exploration and Preprocessing



```
# Calculate the percentage of null values for each column and sort in descend
null_percentage = (df.isnull().sum() / df.shape[0]).sort_values(ascending=Fal
null_percentage
```

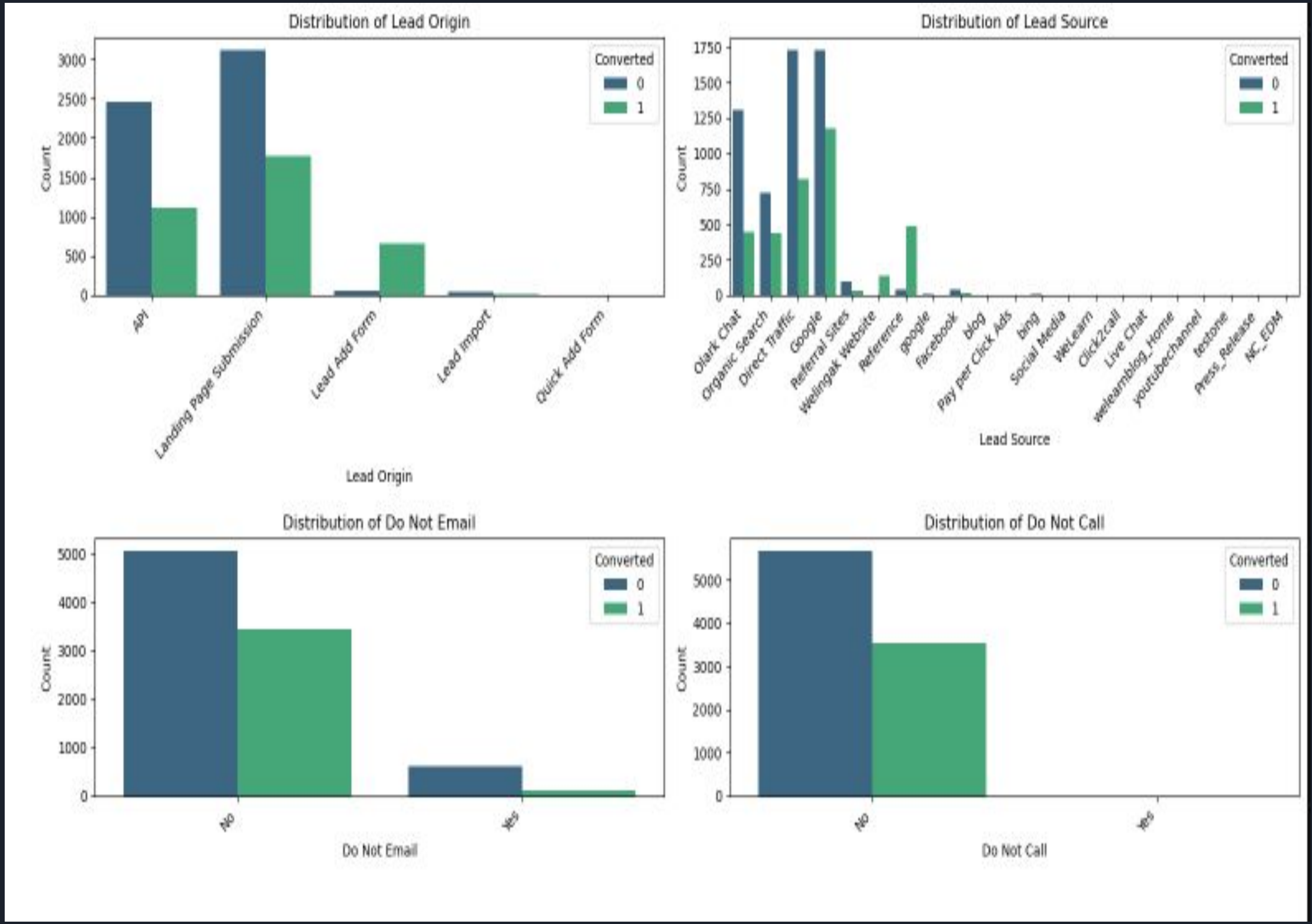
```
Lead Quality                                     51.590909
Asymmetrique Activity Index                      45.649351
Asymmetrique Profile Score                       45.649351
Asymmetrique Activity Score                      45.649351
Asymmetrique Profile Index                       45.649351
Tags                                             36.287879
Lead Profile                                     29.318182
What matters most to you in choosing a course    29.318182
What is your current occupation                  29.112554
Country                                          26.634199
How did you hear about X Education               23.885281
Specialization                                   15.562771
City                                             15.367965
Page Views Per Visit                              1.482684
TotalVisits                                       1.482684
Last Activity                                     1.114719
Lead Source                                       0.389610
Receive More Updates About Our Courses            0.000000
I agree to pay the amount through cheque          0.000000
```

1. **Handling 'Select' Values:**
   - - Replace 'Select' with NaN.

2. **High Cardinality Categorical Variables:**
   - - Drop "Prospect ID" and "Lead Number."

3. **Imputing Missing Values:**
   - Fill "What matters most" and "Current occupation" NaNs with mode.
   - Replace "Country" NaNs with "Unknown."
   - Impute NaNs in "TotalVisits" and "Page Views" with means.
   - Fill "Last Activity" and "Lead Source" NaNs with mode.

Data Exploration and Preprocessing



Count Plots:

1. Each subplot is a count plot for a different categorical variable from the DataFrame.
2. The x-axis represents the unique categories of the respective variable.
3. The y-axis shows the count of occurrences for each category.

Count plots helped us for the identification of patterns, trends, and potential relationships between categorical variables and the target variable ('Converted').

Categorical Variables :

1. Lead Origin_Lead Add Form
2. Lead Source_Olark Chat
3. Lead Source_Welingak Website
4. Do Not Email_Yes
5. Last Notable Activity_Had a Phone Conversation

+ More variables but where removed during vif analysis.

Univariate Analysis & Outlier Analysis

Univariate Analysis Findings:

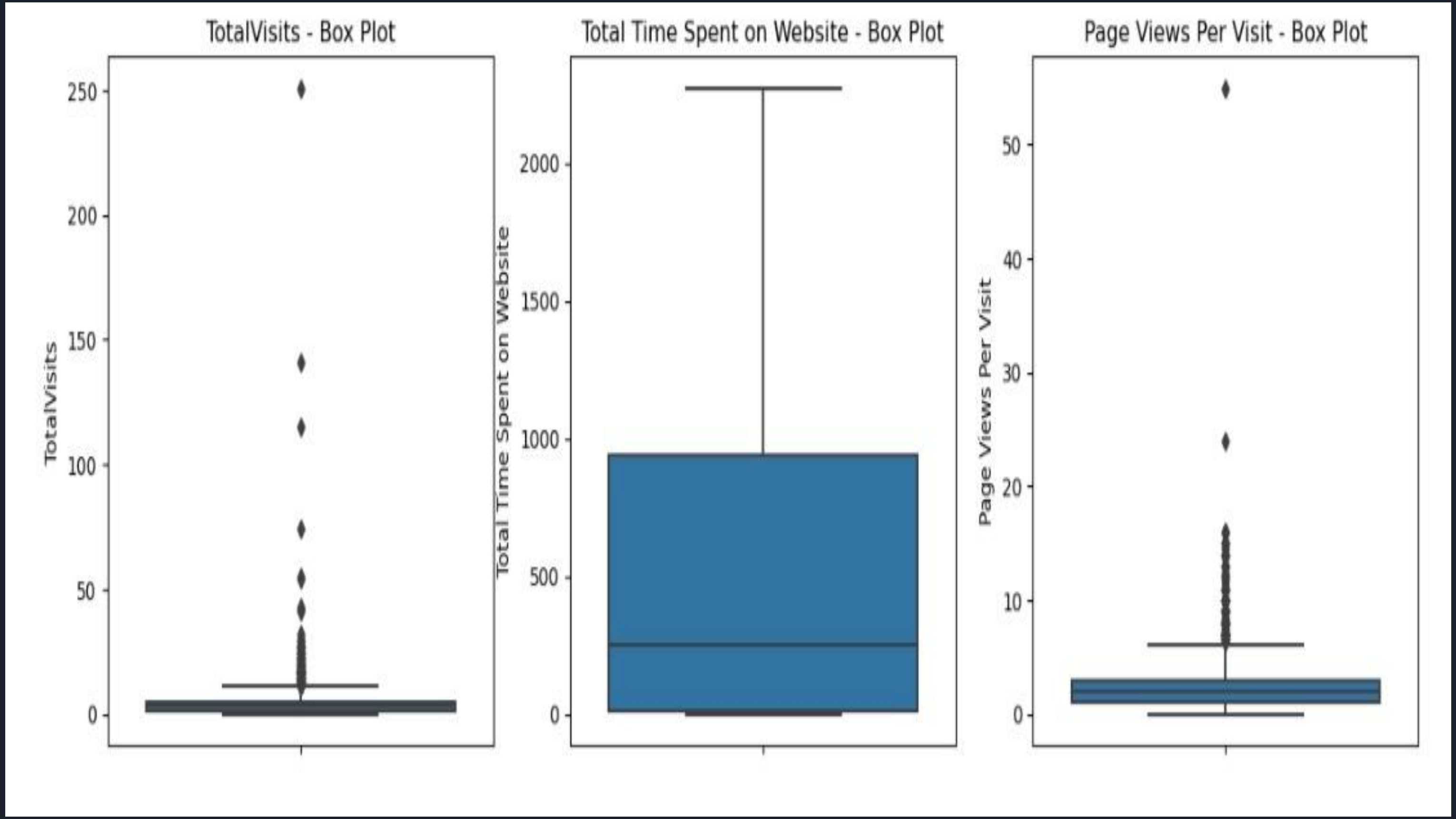
- Identified outliers in key columns through data distribution analysis in the 'Leads' dataset.

Columns with Outliers:

- Outliers detected in:
  - TotalVisits
  - Page Views Per Visit

Outlier Treatment Method:

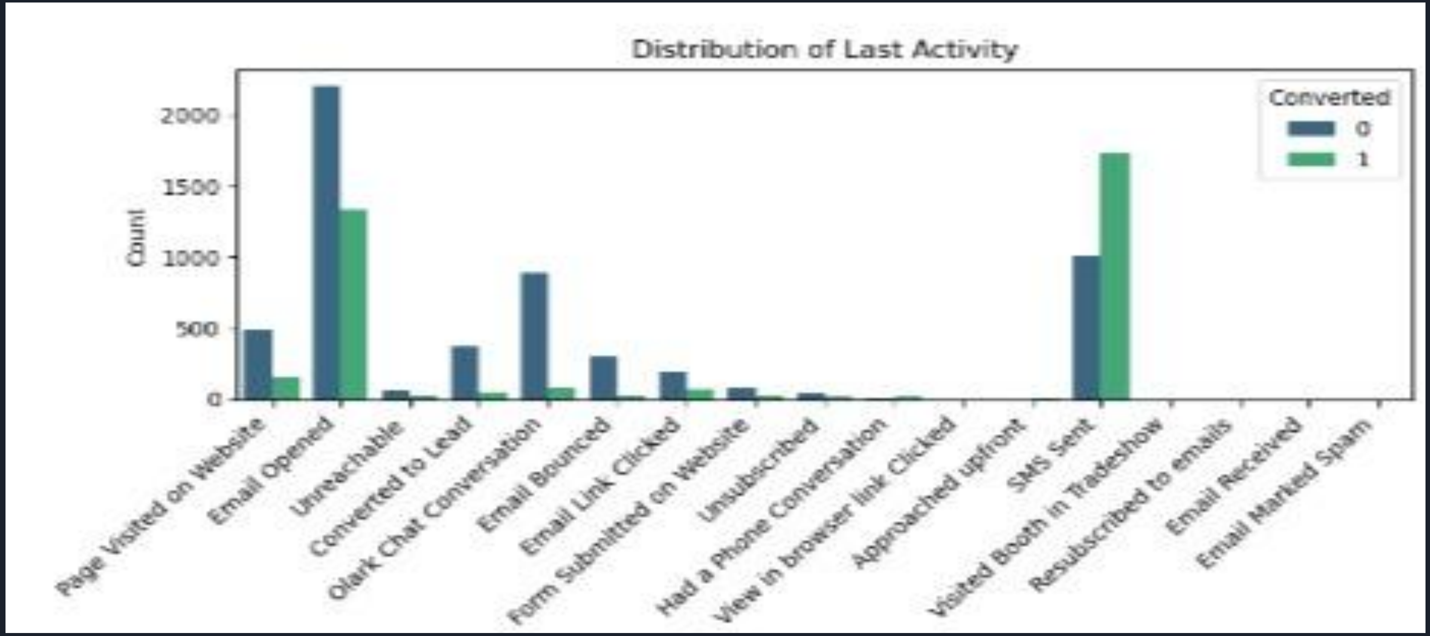
- Opted not to remove any outliers as sometimes people do visit website more than 200 times while purchasing product

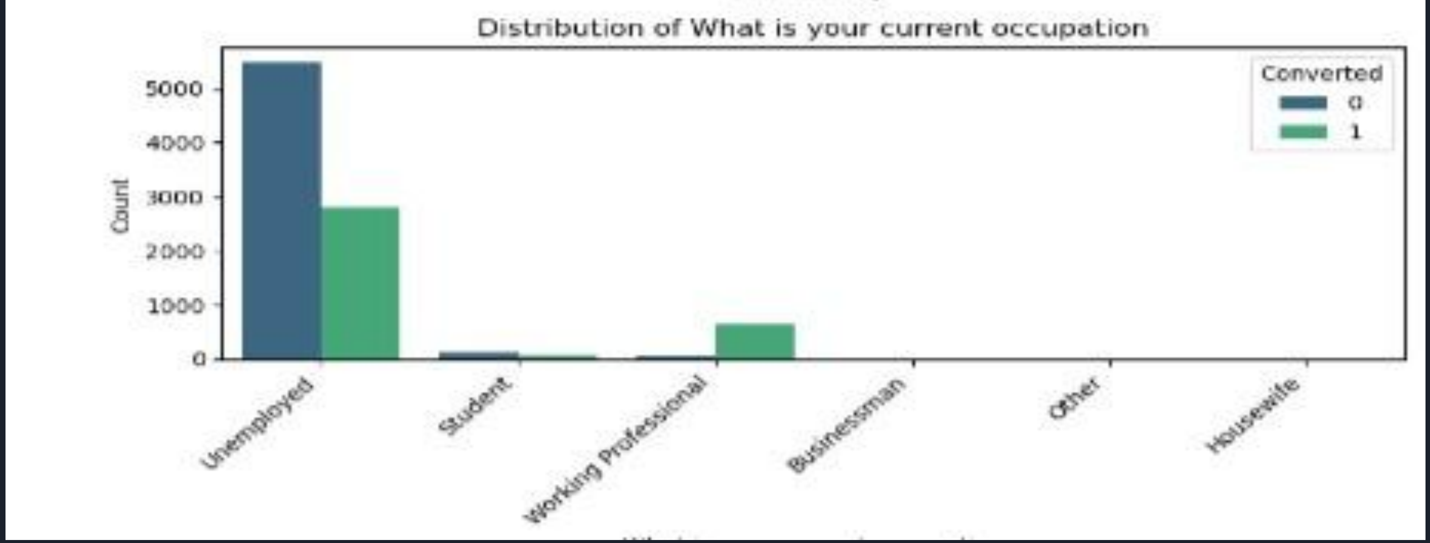


Bivariate Analysis

Bivariate analysis reveals that converted leads are more likely to be:

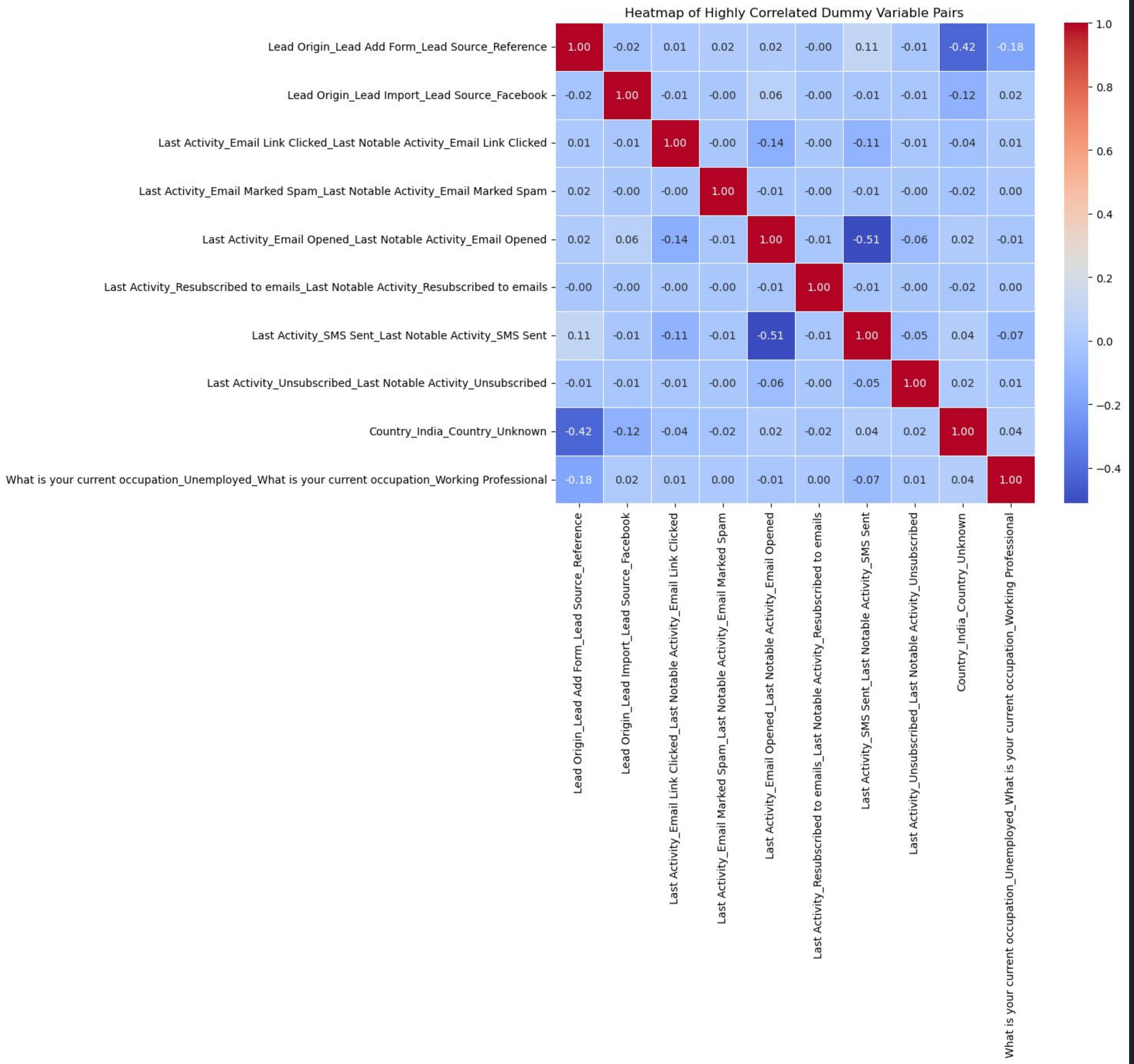
- Lateral students or those expressing interest in the upcoming batch.
- Tagged with "High in Relevance" for lead quality.
- Originated through "Lead Add Form" and "Quick Add Form."
- Associated with specific sources like 'Welingak Website,' 'WeLearn,' 'Live Chat,' and 'NC_EDM.'





High Correlation’s:

- **Dropped Variables:**
  - The following variables as show in the image were dropped
- **Reasons for Dropping:**
  - High correlation with other variables.
  - Redundant information.
  - Limited variability.
- Impact on Analysis:
  - Improved model simplicity.
  - Mitigated multicollinearity.
  - Enhanced interpretability.
- Conclusion:
  - Streamlined dataset for focused analysis.
  - Maintained relevant features for model efficacy.

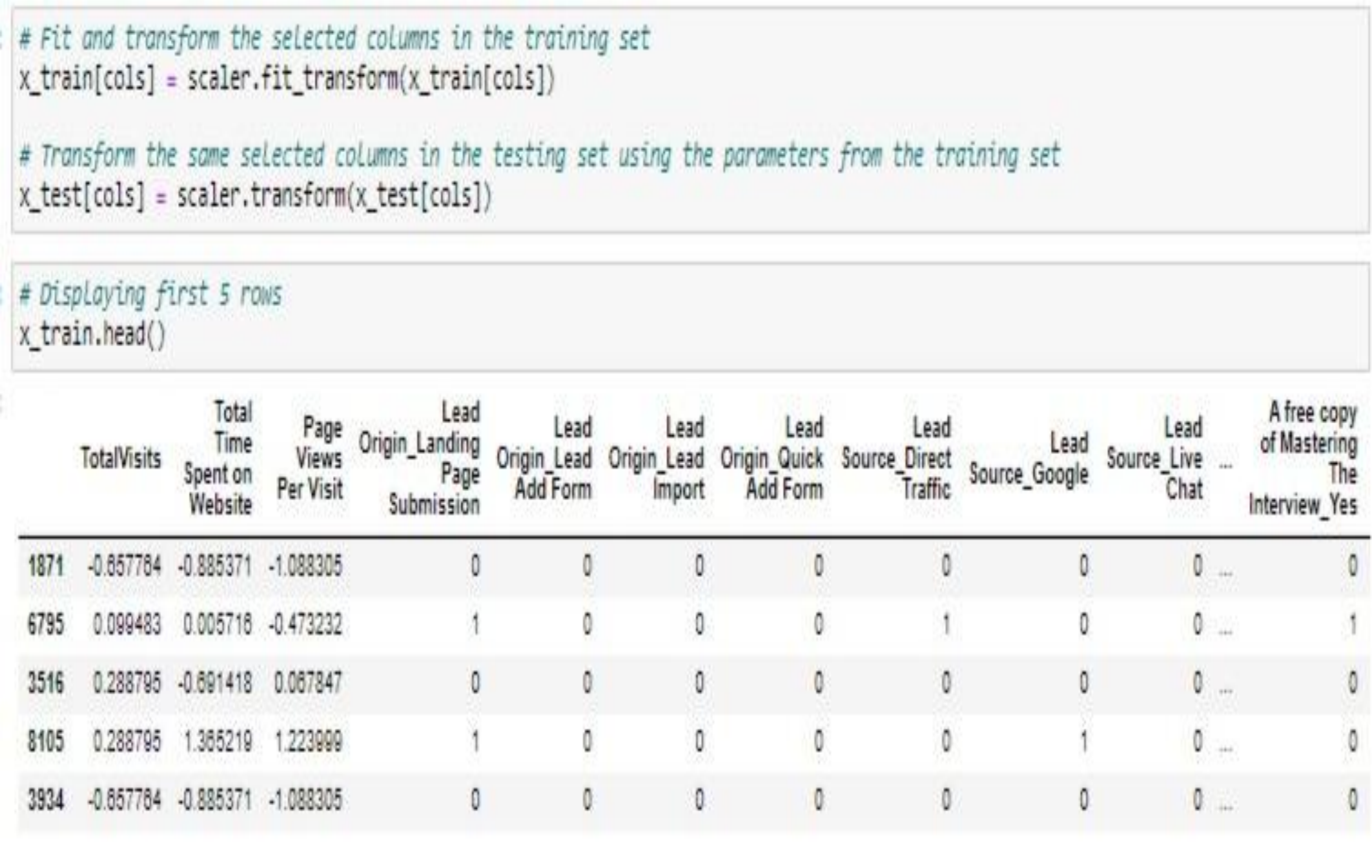


Heatmap of Highly Correlated Dummy Variable Pairs

	Lead Origin_Lead Add Form_Lead Source_Reference	Lead Origin_Lead Import_Lead Source_Facebook	Last Activity_Email Link Clicked_Last Notable Activity_Email Link Clicked	Last Activity_Email Marked Spam_Last Notable Activity_Email Marked Spam	Last Activity_Email Opened_Last Notable Activity_Email Opened	Last Activity_Resubscribed to emails_Last Notable Activity_Resubscribed to emails	Last Activity_SMS Sent_Last Notable Activity_SMS Sent	Last Activity_Unsubscribed_Last Notable Activity_Unsubscribed	Country_India_Country_Unknown	What is your current occupation_Unemployed_What is your current occupation_Working Professional
Lead Origin_Lead Add Form_Lead Source_Reference	1.00	-0.02	0.01	0.02	0.02	-0.00	0.11	-0.01	-0.42	-0.18
Lead Origin_Lead Import_Lead Source_Facebook	-0.02	1.00	-0.01	-0.00	0.06	-0.00	-0.01	-0.01	-0.12	0.02
Last Activity_Email Link Clicked_Last Notable Activity_Email Link Clicked	0.01	-0.01	1.00	-0.00	-0.14	-0.00	-0.11	-0.01	-0.04	0.01
Last Activity_Email Marked Spam_Last Notable Activity_Email Marked Spam	0.02	-0.00	-0.00	1.00	-0.01	-0.00	-0.01	-0.00	-0.02	0.00
Last Activity_Email Opened_Last Notable Activity_Email Opened	0.02	0.06	-0.14	-0.01	1.00	-0.01	-0.51	-0.06	0.02	-0.01
Last Activity_Resubscribed to emails_Last Notable Activity_Resubscribed to emails	-0.00	-0.00	-0.00	-0.00	-0.01	1.00	-0.01	-0.00	-0.02	0.00
Last Activity_SMS Sent_Last Notable Activity_SMS Sent	0.11	-0.01	-0.11	-0.01	-0.51	-0.01	1.00	-0.05	0.04	-0.07
Last Activity_Unsubscribed_Last Notable Activity_Unsubscribed	-0.01	-0.01	-0.01	-0.00	-0.06	-0.00	-0.05	1.00	0.02	0.01
Country_India_Country_Unknown	-0.42	-0.12	-0.04	-0.02	0.02	-0.02	0.04	0.02	1.00	0.04
What is your current occupation_Unemployed_What is your current occupation_Working Professional	-0.18	0.02	0.01	0.00	-0.01	0.00	-0.07	0.01	0.04	1.00

Data Preparation for Modeling

- **Creation of Dummy Variables:**
    - Dummy variables established for independent variables for ease of interpretation and odds ratio calculation.
    - Binary variables encoded with 1 for 'Yes' and 0 for 'No.'
    -
- **Train-Test Split:**
    - 'Leads' dataset split into Train (70%) and Test (30%) sets.
    - Train set utilized for model training, Test set for model evaluation.
    -
- **Feature Scaling:**
    - Ensured uniform scale for all variables to prevent dominance by high-magnitude features.
    - Implemented 'StandardScaler' to standardize data, bringing it to a mean of '0' and standard deviation of '1.'



```
# Fit and transform the selected columns in the training set
x_train[cols] = scaler.fit_transform(x_train[cols])

# Transform the same selected columns in the testing set using the parameters from the training set
x_test[cols] = scaler.transform(x_test[cols])
```

```
# Displaying first 5 rows
x_train.head()
```

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form	Lead Source_Direct Traffic	Lead Source_Google	Lead Source_Live Chat	...	A free copy of Mastering The Interview_Yes
1871	-0.657764	-0.885371	-1.088305	0	0	0	0	0	0	0	...	0
6795	0.099483	0.005716	-0.473232	1	0	0	0	1	0	0	...	1
3516	0.288795	-0.691418	0.067847	0	0	0	0	0	0	0	...	0
8105	0.288795	1.365219	1.223999	1	0	0	0	0	1	0	...	0
3934	-0.657764	-0.885371	-1.088305	0	0	0	0	0	0	0	...	0

Model Building: Using logistic Regression

Model Selection:

- Utilized the Generalized Linear Model (GLM) from the StatsModels library to build the Logistic Regression Model.

Initial Model Features:

- Initially built the model with 93 features from the X_train dataset.

Identifying Insignificant Features:

- Found a majority of features to be insignificant, prompting the need for feature selection.

Feature Selection Technique: Recursive Feature Elimination (RFE)

- Applied RFE, an optimization technique, to identify the best subset of features.
- RFE repeatedly constructs models, selecting the best-performing features based on coefficients.
- Top 20 features were identified for further model building.

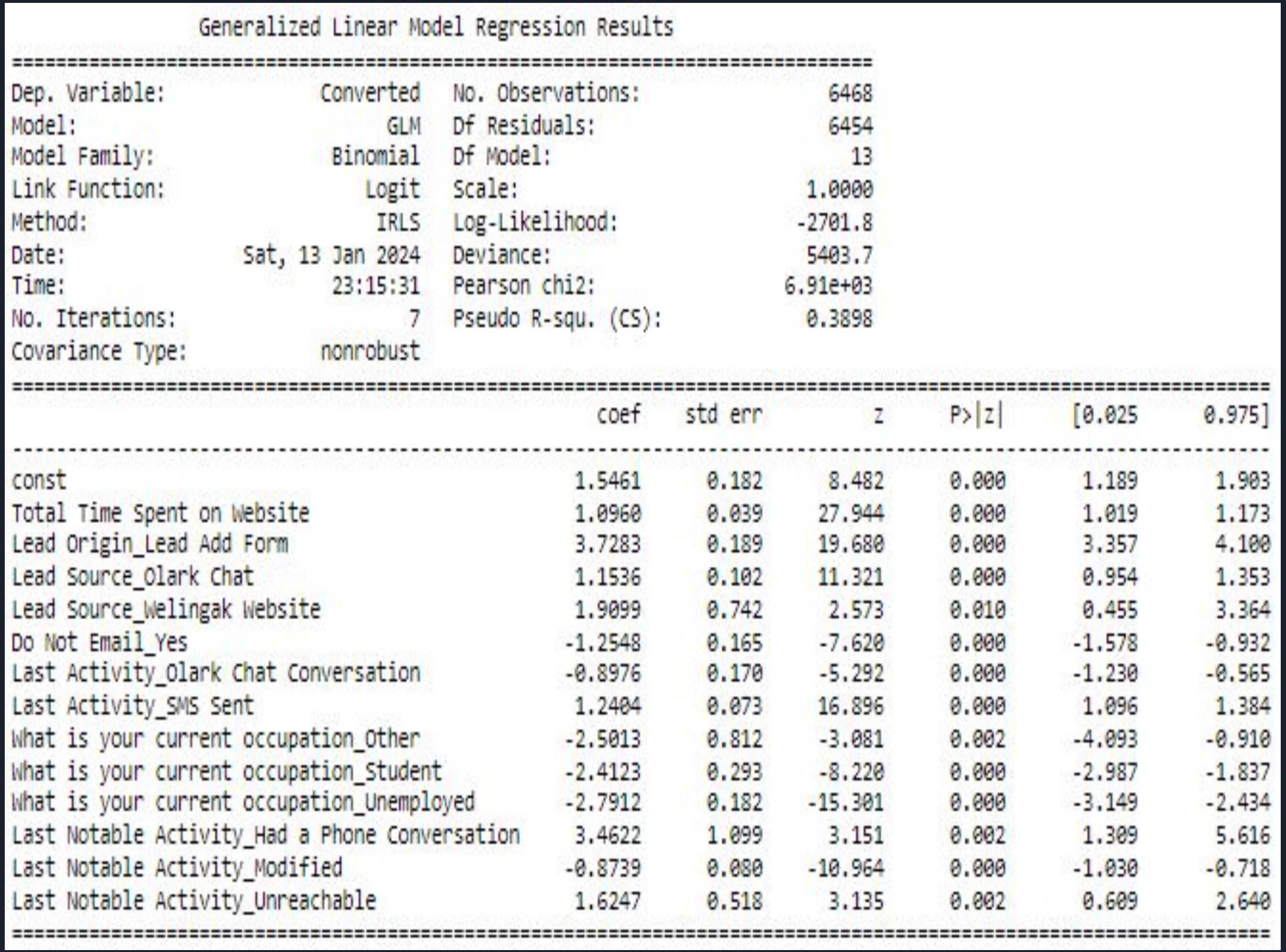
Feature Elimination Criteria:

- Insignificant features were systematically dropped based on P-value and Variance Inflation Factor (VIF).
- Accepted P-value set below 0.05, and VIF kept less than 5 for feature retention.

Iterative Elimination Process:

- Features were eliminated one by one, considering their statistical significance and multicollinearity.

This streamlined approach in feature selection enhances model interpretability and performance by focusing on the most relevant predictors

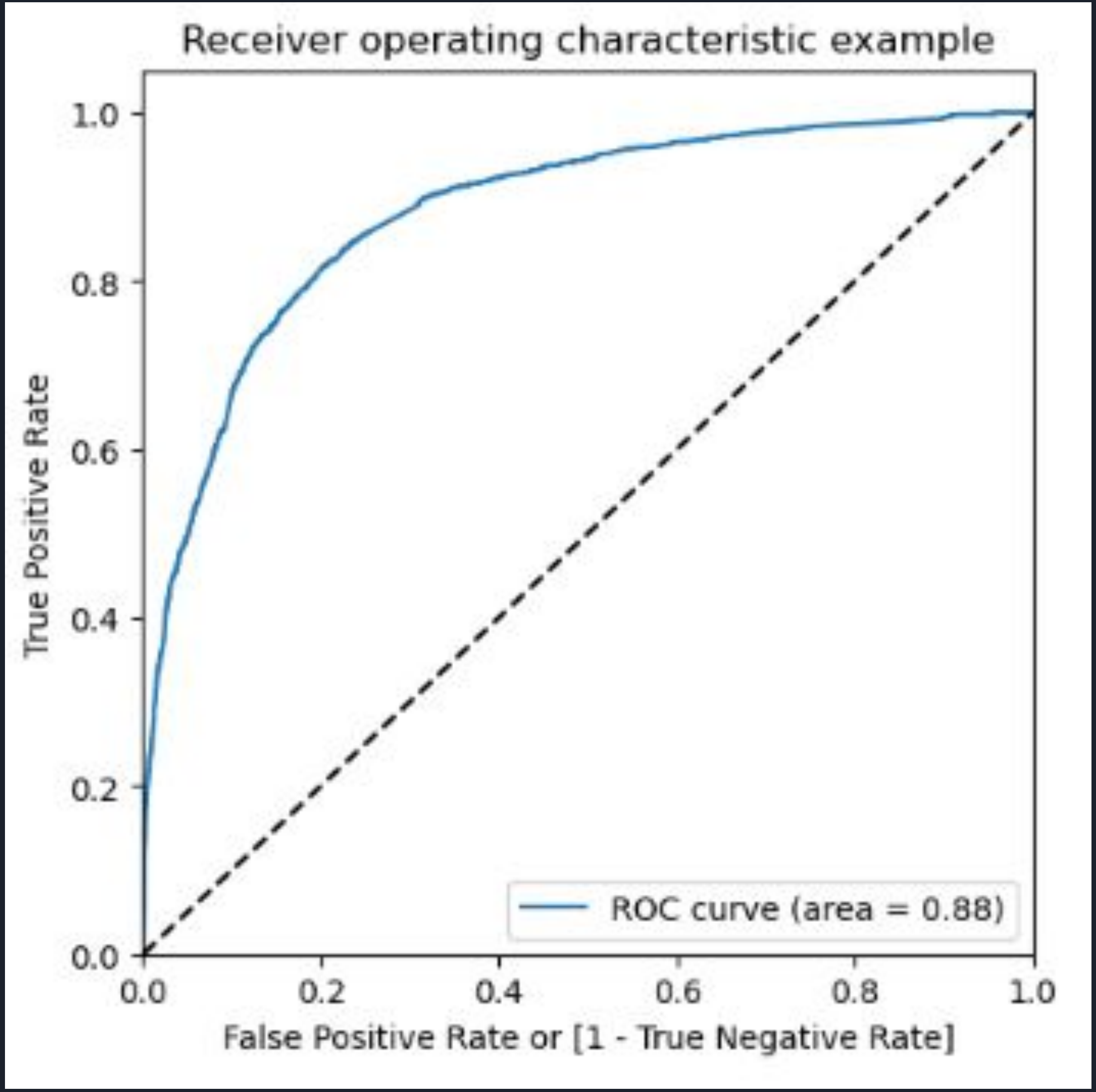


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:             Converted   No. Observations:                 6468
Model:                           GLM   Df Residuals:                     6454
Model Family:               Binomial   Df Model:                           13
Link Function:                 Logit   Scale:                          1.0000
Method:                         IRLS   Log-Likelihood:                -2701.8
Date:               Sat, 13 Jan 2024   Deviance:                       5403.7
Time:                       23:15:31   Pearson chi2:                 6.91e+03
No. Iterations:                    7   Pseudo R-squ. (CS):             0.3898
Covariance Type:           nonrobust
===================================================================================================================
                                                      coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------------
const                                               1.5461      0.182      8.482      0.000       1.189       1.903
Total Time Spent on Website                         1.0960      0.039     27.944      0.000       1.019       1.173
Lead Origin_Lead Add Form                           3.7283      0.189     19.680      0.000       3.357       4.100
Lead Source_Olark Chat                              1.1536      0.102     11.321      0.000       0.954       1.353
Lead Source_Welingak Website                        1.9099      0.742      2.573      0.010       0.455       3.364
Do Not Email_Yes                                   -1.2548      0.165     -7.620      0.000      -1.578      -0.932
Last Activity_Olark Chat Conversation              -0.8976      0.170     -5.292      0.000      -1.230      -0.565
Last Activity_SMS Sent                              1.2404      0.073     16.896      0.000       1.096       1.384
What is your current occupation_Other              -2.5013      0.812     -3.081      0.002      -4.093      -0.910
What is your current occupation_Student            -2.4123      0.293     -8.220      0.000      -2.987      -1.837
What is your current occupation_Unemployed         -2.7912      0.182    -15.301      0.000      -3.149      -2.434
Last Notable Activity_Had a Phone Conversation      3.4622      1.099      3.151      0.002       1.309       5.616
Last Notable Activity_Modified                     -0.8739      0.080    -10.964      0.000      -1.030      -0.718
Last Notable Activity_Unreachable                   1.6247      0.518      3.135      0.002       0.609       2.640
===================================================================================================================
```

Final Model Interpretation

Receiver Operating Characteristics (ROC) Curve:

- **AUC Assessment:**
  - The Area Under the Curve (AUC) of the ROC curve is a pivotal metric for evaluating the model's performance.
- **Goodness of the Model:**
  - The ROC curve's proximity to the upper-left section of the graph signifies the effectiveness of the model.
- **Model Evaluation:**
  - With an AUC value of 0.88, our model demonstrates a high level of discriminative ability and accuracy in distinguishing between positive and negative instances.



Evaluation Metric

Feature Selection:

- The final model includes 13 key features that satisfy selection criteria, enhancing model efficiency and interpretability.

Prediction Threshold:

- Lead scores with a conversion probability greater than 0.35 are classified as "Converted."

Test Dataset Prediction:

- Utilized the 0.35 probability threshold to predict conversions for leads in the test dataset.

Confusion Matrix (cut-off 0.35):

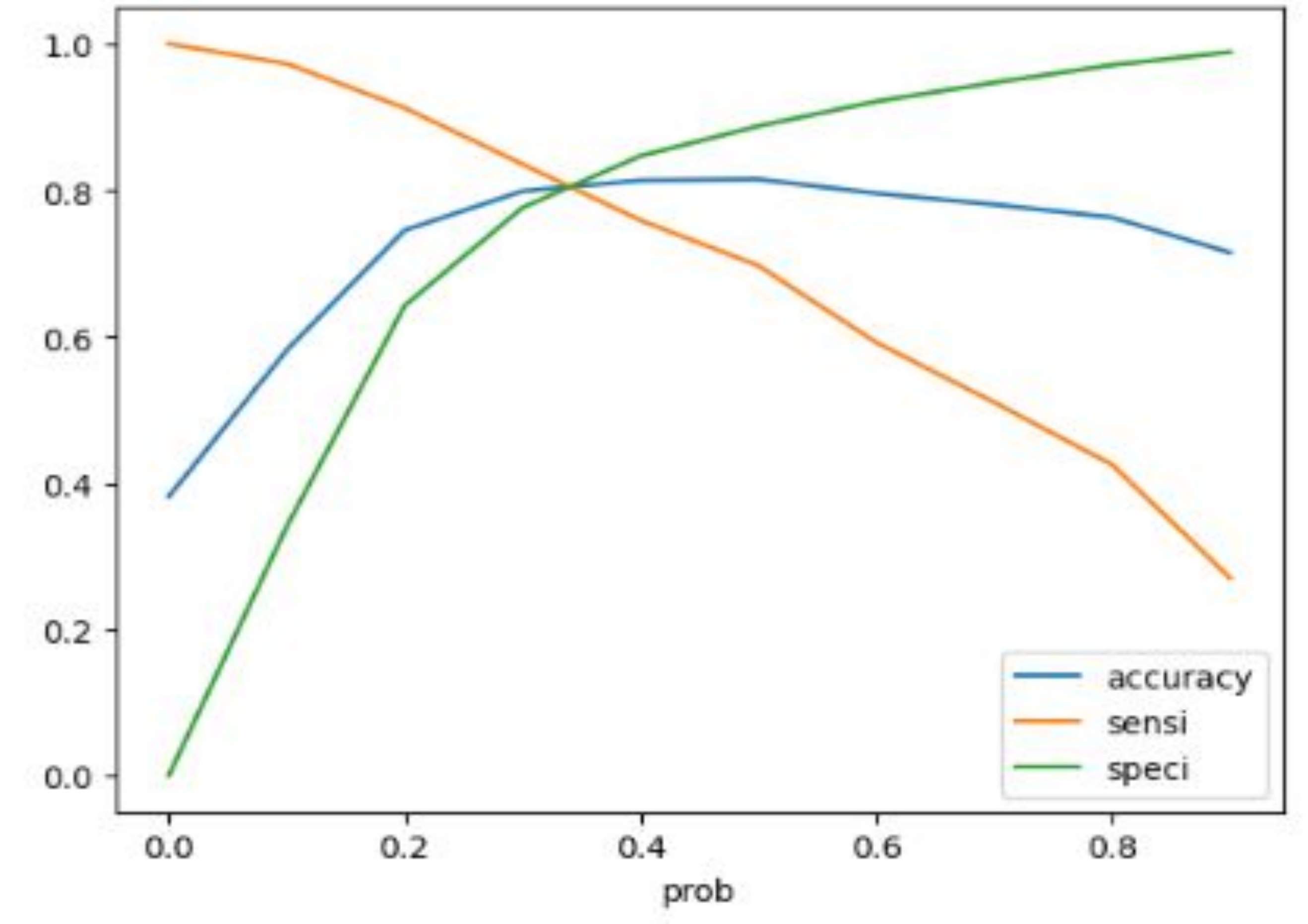
- [[3272 , 730]
- [514 , 1952]]

Evaluation Metrics:

- Accuracy: 0.8076
- Sensitivity (Recall): 0.7915
- Specificity: 0.8175

Sensitivity gauges the model's accuracy in correctly identifying actual positive instances, specifically converted leads.

Rationale: The emphasis on maximizing sensitivity is rooted in the goal of capturing a larger proportion of leads that have the potential to convert. By doing so, the model minimizes missed opportunities, ensuring a more comprehensive identification of positive outcomes.



From the curve above, 0.35 is the optimum point to take it as a cutoff probability.

Evaluation Metric : Precision & Recall Tradeoff

Feature Selection:

- The final model includes 13 key features that satisfy selection criteria, enhancing model efficiency and interpretability.

Prediction Threshold:

- Lead scores with a conversion probability greater than 0.40 are classified as "Converted." using

Test Dataset Prediction:

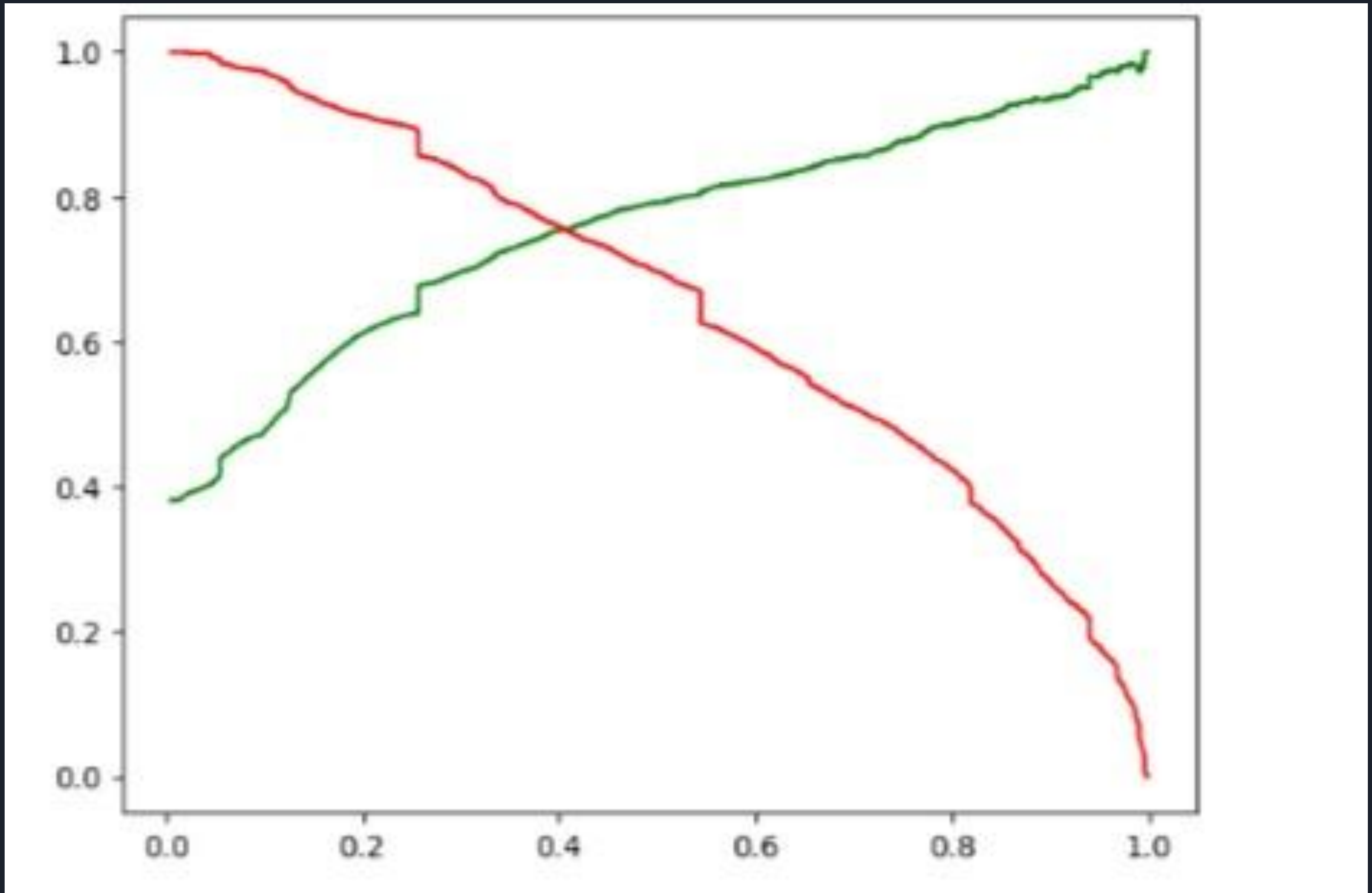
- Utilized the 0.40 probability threshold to predict conversions for leads in the test dataset.

Confusion Matrix (cut-off 0.40):

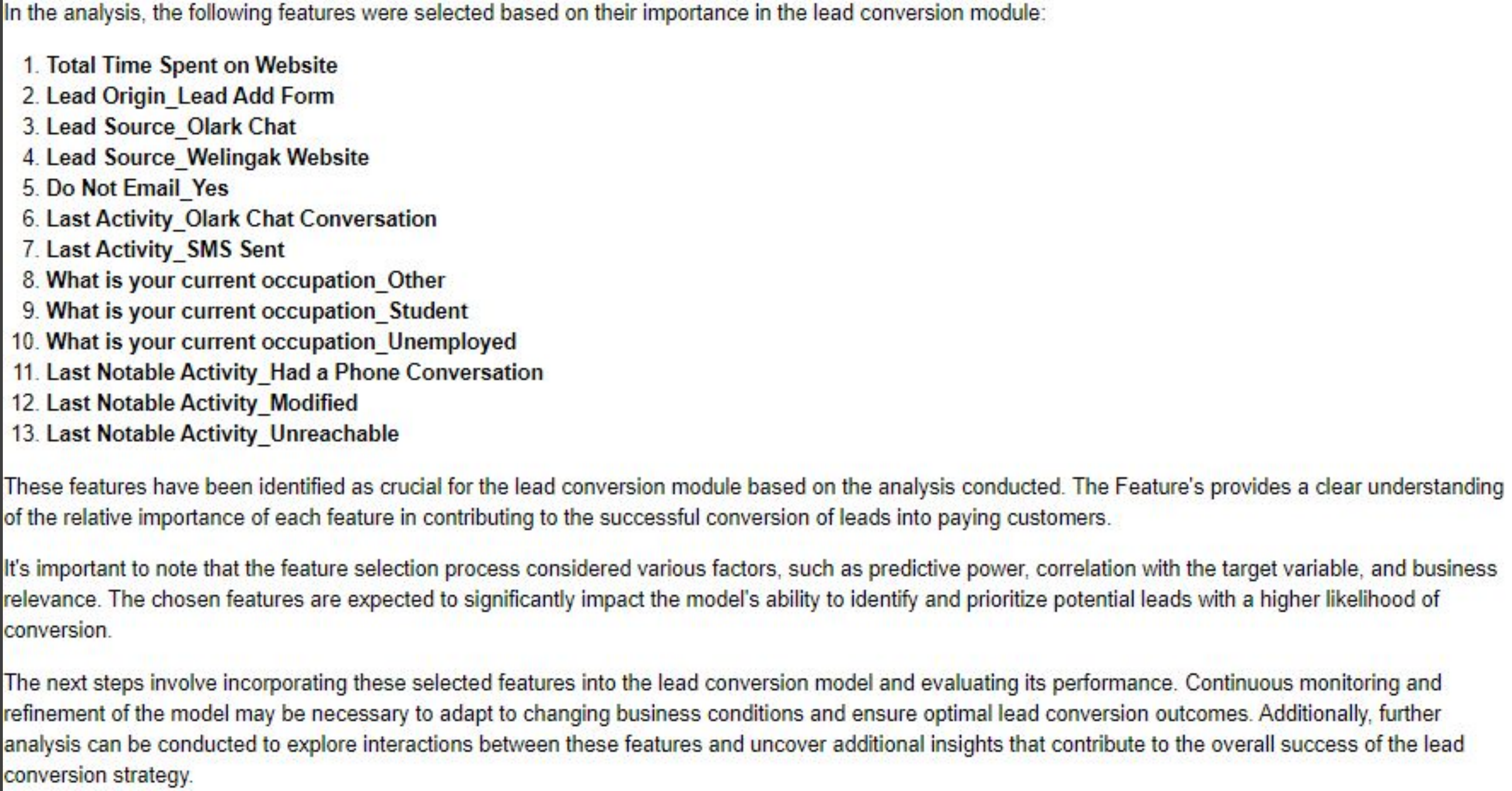
- [[1433 , 244]
- [260 , 835]]

Evaluation Metrics:

- Precision: 0.7278
- Accuracy : 0.8181
- Sensitivity (Recall): 0.7915
- Specificity: 0.8545



Feature Selection :



In the analysis, the following features were selected based on their importance in the lead conversion module:

1. **Total Time Spent on Website**
2. **Lead Origin_Lead Add Form**
3. **Lead Source_Olark Chat**
4. **Lead Source_Welingak Website**
5. **Do Not Email_Yes**
6. **Last Activity_Olark Chat Conversation**
7. **Last Activity_SMS Sent**
8. **What is your current occupation_Other**
9. **What is your current occupation_Student**
10. **What is your current occupation_Unemployed**
11. **Last Notable Activity_Had a Phone Conversation**
12. **Last Notable Activity_Modified**
13. **Last Notable Activity_Unreachable**

These features have been identified as crucial for the lead conversion module based on the analysis conducted. The Feature's provides a clear understanding of the relative importance of each feature in contributing to the successful conversion of leads into paying customers.

It's important to note that the feature selection process considered various factors, such as predictive power, correlation with the target variable, and business relevance. The chosen features are expected to significantly impact the model's ability to identify and prioritize potential leads with a higher likelihood of conversion.

The next steps involve incorporating these selected features into the lead conversion model and evaluating its performance. Continuous monitoring and refinement of the model may be necessary to adapt to changing business conditions and ensure optimal lead conversion outcomes. Additionally, further analysis can be conducted to explore interactions between these features and uncover additional insights that contribute to the overall success of the lead conversion strategy.

Conclusion and Recommendations:

- **Lead Origin - Lead Add Form (Coefficient: 3.7283):**
  - Leads originating from the "Lead Add Form" have a substantial positive impact on conversion, with a coefficient of 3.7283.
- **Last Notable Activity - Had a Phone Conversation (Coefficient: 3.4622):**
  - Leads with the last notable activity being a "Phone Conversation" significantly contribute to conversion, with a coefficient of 3.4622.
- **Lead Source - Welingak Website (Coefficient: 1.9099):**
  - The source of the lead being the "Welingak Website" is another crucial factor influencing conversion, with a coefficient of 1.9099.

t

- S**horter Sales Cycle:**
  - Intuitive prioritization ensures a more rapid progression through the sales cycle.
- **Improved Opportunity-to-Deal Ratio:**
  - Concentrating efforts on hot leads increases the likelihood of converting opportunities into successful deals.
- **Control Over Volatile Buying Cycle:**
  - Targeting hot leads provides a more controlled approach to navigate the unpredictable nature of the buying cycle.
- **Increased Marketing Effectiveness:**
  - Precision targeting enhances the efficiency and effectiveness of marketing strategies.
- **Enhanced Sales Forecasting:**
  - The model's identification of hot leads contributes to more accurate sales forecasting and strategic planning.
- **Minimized Opportunity Loss:**
  - Focusing on leads with a higher probability of conversion minimizes the loss of potential opportunities.
- **Revenue Boost:**
  - The overall impact translates into an increase in revenue, driven by a more strategic and targeted approach.